PAT-NO: JP403225445A

DOCUMENT-IDENTIFIER: JP 03225445 A

TITLE: LOAD DISPERSION STRIPING SYSTEM

PUBN-DATE: October 4, 1991

INVENTOR-INFORMATION:

NAME	COUNTRY
NAKA, SEIICHIRO	
YASUGADAIRA, MICHIKO	

ASSIGNEE-INFORMATION:

NAME	COUNTRY
NEC CORP	N/A
TOHOKU NIPPON DENKI SOFTWARE KK	N/A

APPL-NO: JP02021341

APPL-DATE: January 30, 1990

INT-CL (IPC): G06F012/00 , G06F003/06

ABSTRACT:

PURPOSE: To prevent input and output from being congested on a specific storage device by determining storage devices to be assigned in predetermined order and updating a storage device control table according to the number of storage device and an assignment quantity which are calculated from the states of the respective storage devices stored in a storage device control table.

CONSTITUTION: The storage device control table 25 is referred to pieces of information on the input and output of the respective storage devices 4 to select a necessary number of storage devices in order from the device having the least input/output request quantity; when there are devices having the same quantity, the lengths of their queues are compared to select a shorter-length storage device and then a storage device having the largest data storage capacity.

Further, the storage device having the largest data storage capacity is selected in the storage device control table 25 and after the storage devices to be stored with data are determined, the data storage capacity of each storage device is updated. Actual input and output operation is performed by an input/output control part 3 and the length of the input/output queue and input/ output request quantity in the referred storage device control table 25 are updated. Consequently, a load is prevented from being concentrated on a specific storage device.

⑲日本国特許庁(JP)　⑪特許出願公開

⑫公開特許公報(A)　平3−225445

⑤1Int.Cl.⁵	識別記号	庁内整理番号
G 06 F 12/00	301 B	8944−5B
3/06	301 J	7232−5B

⑬公開　平成3年(1991)10月4日

審査請求　未請求　請求項の数 1　(全6頁)

⑤4発明の名称　負荷分散ストライピング方式

⑳特　願　平2−21341
㉒出　願　平2(1990)1月30日

⑦2発明者　中　誠一郎　東京都港区芝5丁目33番1号　日本電気株式会社内
⑦2発明者　安ケ平　達子　宮城県仙台市青葉区中央4丁目6番1号　東北日本電気ソフトウエア株式会社内
⑦1出願人　日本電気株式会社　東京都港区芝5丁目7番1号
⑦1出願人　東北日本電気ソフトウエア株式会社　宮城県仙台市青葉区中央4丁目6番1号
⑦4代理人　弁理士　内原　晋

明　細　書

発明の名称

負荷分散ストライピング方式

特許請求の範囲

入出力を制御する機構を備えたファイルシステムが、同一な入出力の処理速度を持つ複数の記憶装置上に、同一容量で分割して格納されている仮想パーティション方式における負荷分散ストライピング方式において、前記記憶装置の転送速度、処理効率を保つために必要なデータ量、要求される処理速度、記憶装置数を格納している仮想パーティション管理テーブルと、割り当て処理の際に、要求されているデータ割り当て量と前記仮想パーティション管理テーブルに格納された要求される処理速度から、前記要求される処理速度を満足するようにデータを分散させる記憶装置数と割り当て量を算出する手段と、各記憶装置の状態を示す格納できるデータ量、入出力に対する待ち行列の長さ、入出力要求量とが格納されている記憶装置管理テーブルと、前記記憶装置管理テーブルに格納された前記各記憶装置の状態から前記算出された記憶装置数および割り当て量に従って予め定めた順に割り当てる記憶装置を決定するとともに前記記憶装置管理テーブルを更新する手段とを有することを特徴とする負荷分散ストライピング方式。

発明の詳細な説明

〔産業上の利用分野〕

本発明は、複数の記憶装置を資源とするコンピュータシステムのファイルシステムに関する。

〔従来の技術〕

従来のデータ格納方式は、データ量が多く一つの記憶装置に格納できない場合に、格納限度を越えた分を次の記憶装置に格納するような方式、または割り当てを行う各記憶装置の情報をシステムのテーブルに格納しておき、その情報を基に記憶装置を選択割り当てを行う方式、ユーザプログラ

ムでファイルを作成する際に使用する記憶装置を指定するような方式をとっていた。

〔発明が解決しようとする課題〕

上記従来のデータ格納方式では、ある特定の記憶装置にのみ入出力が集中したり、その反面で、全く入出力要求がない記憶装置が存在してしまうなど、特定の記憶装置に対する負荷が増加し、処理速度が低下するという問題点、また、ファイル作成時にストライピングの数が固定され、ユーザに物理的な記憶装置を意識させてしまうといった問題点がある。

〔課題を解決するための手段〕

本発明の負荷分散ストライピング方式は、各記憶装置の転送速度、処理効率を保つために必要なデータ量、要求される処理速度、記憶装置数を格納している仮想パーティション管理テーブルと、割り当て処理の際に、要求されているデータ割り当て量と前記仮想パーティション管理テーブルに格納された要求される処理速度から、前記要求される処理速度を満足するようにデータを分散させる記憶装置数と割り当て量を算出する手段と、各記憶装置の状態を示す格納できるデータ量、入出力に対する待ち行列の長さ、入出力要求量とが格納されている記憶装置管理テーブルと、前記記憶装置管理テーブルに格納された前記各記憶装置の状態から前記算出された記憶装置数および割り当て量に従って予じめ定めた順に割り当てる記憶装置を決定するとともに前記記憶装置管理テーブルを更新する手段とを有している。

〔実施例〕

次に本発明について図面を参照して説明する。

第１図は、本発明の構成を表す図である。

本発明は、ユーザプログラム１からの入出力に関する処理について、同一の処理速度と容量を持つｎ個からなる記憶装置４（ＤＫ１、ＤＫ２、…、ＤＫｎ）と、それら記憶装置を制御する部分からなる。記憶装置４には既に仮想パーティション方式で生成されたファイルがパーティションＶＰとして格納されている。制御する部分はファイルシステム２としてファイル管理装置２と、データ解放制御部２２と、入出力制御部３と、データの割り当てを制御するデータ割当制御手段２３と、仮想パーティション管理テーブル２４と、記憶装置管理テーブル２５で構成されている。

データ割当制御部２３は、分散格納する記憶装置数と割り当て量を算出する手段２３１と、データを格納する記憶装置を決定する手段２３２とによって構成される。

第２図は、仮想パーティション管理テーブル２４を示す概念図であり、パーティションＶＰの転送速度Ｔ、要求処理速度Ｖ、記憶装置の処理効率を保つために必要なデータ処理量Ｕ、記憶装置数Ｎが格納されている。

第３図は、記憶装置管理テーブル２５を示す概念図であり、第２図のパーティションＶＰに対応する。各記憶装置ＤＫ１～ＤＫｎの格納可能なデータ量Ｍ１～Ｍｎ、入出力の待ち行列の長さＱ１～Ｑｎ、入出力の待ち行列全体のデータ量Ｌ１～Ｌｎが格納されている。ここで、入出力の待ち行列の長さと入出力の待ち行列全体のデータ量は入出力制御部３において、入出力要求発生時に更新されているものとする。

仮想パーティション方式によって構築されたファイルシステムＶＰに、ユーザプログラム１からＸバイトの書き込み要求が発生する。ユーザプログラム１からの命令は、ファイル管理装置２１でＲＥＡＤ制御部２２３、ＷＲＩＴＥ制御部２２１、ＤＥＬＥＴＥ制御部２２２に分けられ、それぞれの実行機構に制御が移る。

Ｘバイトの書き込み要求は、ＷＲＩＴＥ制御部２２１を経て、データ格納領域のデータ割当制御部２３に制御が移される。データ割当制御部２３における記憶装置数・割り当て量算出手段２３１では、書き込み要求に伴い発生するデータ割り当てが要求されると、パーティションＶＰに対応する仮想パーティション管理テーブル２４を参照しながら、仮想パーティションに要求されている処理速度を満足するように分割するデータ量と記憶装置の数を算出する。

要求データXバイトを分割する際の手順は、次の通りである。

まず、仮想パーティション管理テーブル２４よりパーティションＶＰの実際の記憶装置の転送速度がＴ、要求されている処理速度はＶ、記憶装置の処理効率を保つために必要なデータ処理量がＵであることが分かる。

ここで、α×Ｕバイト（α＞０　整数）をα台の記憶装置にＵバイトずつ同時にアクセスした場合、その転送速度Ｖ（α）は、

$$V(\alpha)=\alpha\times T \qquad (式1)$$

で表すことができる。

よって、要求されたＸバイトをｑ個の記憶装置にｐ×Ｕバイトずつ平均的に分割格納するとすると、次の関係が成り立たなければならない。

$$X\leq p\times U\times q \qquad (式2)$$

$$V\leq q\times T \qquad (式3)$$

ただし、ｐ≧１（整数）、１≦ｑ≦ｎ（整数）とする。この関係式より、

$$q\geq V/T \qquad (式4)$$

$$p\geq X/(q\times U) \qquad (式5)$$

が成り立つ。

式４を満足する最小の記憶装置数ｑをq_1で表すと

$$q_1=[V/T] \qquad (式6)$$

（［　］はガウス記号である。）

このq_1を基に式５は次のように変形される。

$$p\geq X/(q_1\times U) \qquad (式7)$$

式７を満足する最小のデータ割り当て量ｐをp_1で表すと、

$$p_1=[X/(q_1\times U)] \qquad (式8)$$

また、

$$p_1\geq X/(q_1\times U) \qquad (式9)$$

ゆえに、

$$q_1\geq X/(p_1\times U) \qquad (式10)$$

よって、実際に割り当てる記憶装置数をq_2とすると、

$$q_2=[X/(p_1\times U)]\leq q_1 \qquad (式11)$$

で求めることができる。

ただし、Ｘ＜Ｕの場合は、性能低下を避けるために常に１つの記憶装置にＸバイト全てを割り当てることとする。

このようにして、Ｘバイトをq''個の記憶装置に

$$(p_1\times U,\ p_1\times U,\ \cdots,\ X-(p_1\times U\times(q_2-1)))$$

で分割する量が算出される。

次に、記憶装置選択手段２３２について説明する。

記憶装置数・割り当て量算出手段２３１より、q_2個の記憶装置に分割することを決定したが、

$$n\subset q_2 \qquad (式12)$$

で算出することができる。

本発明においては、各記憶装置に対する負荷をできるだけ均等化するために、記憶装置管理テーブル２５より、各々の記憶装置の入出力に対する情報を参照し、入出力要求量の最も少ないものから順にq_2個の記憶装置を選択することとする。同一のものがあった場合は、待ち行列の長さを比較し、その短いもの、次に格納可能なデータ容量の最も大きい記憶装置から選択することとする。記憶装置管理テーブル２５より格納可能なデータ容量の最も大きい記憶装置から選択することとする。データを格納する記憶装置が決定した後で、各記憶装置のデータ格納可能量の更新を行う。実際の入出力は入出力制御部３で行われ、前記参照した記憶装置管理テーブル２５の入出力待ち行列の長さと入出力要求量が更新される。

次にＸバイトのデータ解放要求がユーザプログラム１から発生した場合について述べる。

この場合は、ファイル管理装置２１のＤＥＬＥＴＥ制御部２２２を経て、データ解放制御部２２に制御が移る。ここで、各記憶装置に対応する記憶装置管理テーブル２５のデータ格納可能量の更新を行い、入出力制御部３で実際のデータ領域の解放処理が行われる。

最後に、Ｘバイトのデータ読み込み要求がユーザプログラム１から発生した場合について述べる。

ファイル管理装置２１のＲＥＡＤ制御部２３３を経て、入出力制御部３において、実際のデータ読み込みが行われるが、ここで、データが割り付けられている各記憶装置に対応する記憶装置管理テーブル２５の入出力待ち行列の長さ、入出力要求量が更新される。

次に、ここまでに述べた負荷分散パーティション方式について、具体的な数字をあげてその処理を説明する。

第４図、第５図は仮想パーティションＶＰに対する仮想パーティション管理テーブル２４と記憶装置管理テーブル２５を示す概念図である。

仮想パーティションＶＰは、４台の記憶装置（ＤＫ１，ＤＫ２，ＤＫ３，ＤＫ４）で構成されており、その処理速度は１００で皆同一である。この性能を満足するために必要なデータ量が２０であり、ユーザから要求されている処理速度は１５０である。この環境の中に、ユーザより１４０のデータ割り当て要求が発生したとする。

まず、記憶装置数・割り当て量案出手段２３１により、割り当てるべき装置台数q_1及びその割り当て量p_1を計算する。式６、式８より、

$q_1 = [150 / 100] = 2$

$p_1 = [140 / (2 \times 20)] = 4$

よって、式１２より実際に割り当てる台数q_2は、

$q_2 = [140 / (3 \times 20)] = 2$

以上より、要求量１４０は２台の記憶装置に（８０，６０）と分散させることとする。

次に、第５図より各記憶装置の入出力負荷状態を参照して、２台の記憶装置を決定する。

組合せとしては装置が４台あるので、$4C2 = 6$通り考えられるが、まず、入出力要求量（必要格納量２０を単位として表す）の最も少ない順で考えると、

ＤＫ１＜ＤＫ２＝ＤＫ３＜ＤＫ４

しかし、ＤＫ２とＤＫ３の入出力要求量が同一であるので、次に、入出力待ち行列の長さを参照すると、

ＤＫ１＜ＤＫ３＜ＤＫ２＜ＤＫ４

よって、ＤＫ１とＤＫ３を２台の記憶装置として選択することになる。

したがって、この例の場合、要求量１４０は次のようにストライピングは、

（ＤＫ１，ＤＫ２，ＤＫ３，ＤＫ４）

－（６０，０，６０，０）

この分散を施したあとの記憶装置テーブルを第６図に示す。

前記データ割り当てを実施した後で、ユーザより１００のデータ解放要求が発生したとすると、ファイル管理装置２１において、実際データが割り付けられている記憶装置を得る。ここでは、（ＤＫ２，ＤＫ４）＝（６０，４０）の割合で割り付けられていたと仮定する。

データ解放要求はＤＥＬＥＴＥ制御部２２２を経て、データ解放制御部２２に制御が移り、第６図で表される記憶装置管理テーブル２５の格納可能量を更新する。

ＤＫ２…４８０→４８０＋６０＝５４０

ＤＫ４…５００→５００＋４０＝５４０

この解放を施したあとの記憶装置管理テーブル２５を第７図に示す。

〔発明の効果〕

以上説明したように本発明は、ユーザに物理的な記憶装置を意識させずに、ファイル書き込み時に動的にストライピングの数を決定し、実際の入出力の負荷状態を基に、ユーザが求める入出力の性能を実現し、特定の記憶装置に入出力が集中することを防ぐよう効率的な記憶装置の活用とデータの割り当てを行う効果がある。

図面の簡単な説明

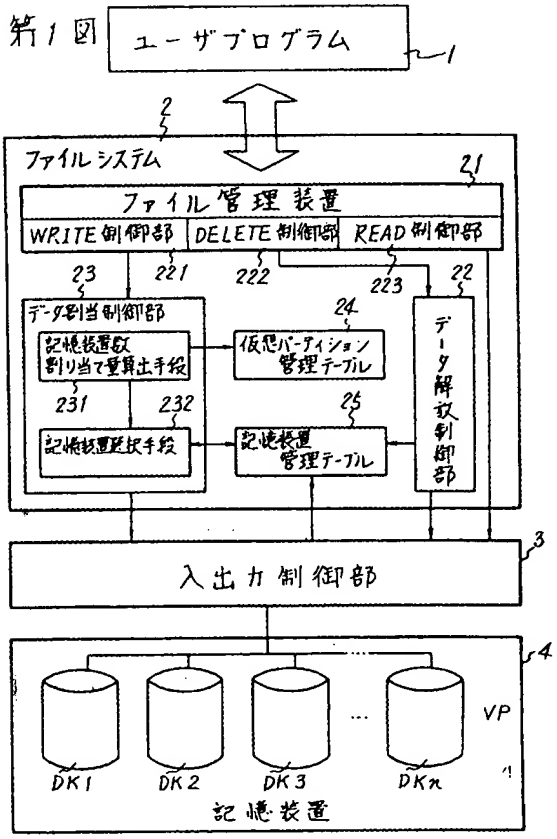
第１図は本発明の一実施例を示す図、第２図は第１図に示す仮想パーティション管理テーブルを示す図、第３図は第１図に示す記憶装置管理テーブルを示す図、第４図は具体的な数字を使って仮想パーティション管理テーブルを示した図、第５図は具体的な数字を使って記憶装置管理テーブルを示した図、第６図は分散を実施したあと更新した記憶装置管理テーブルを示した図、第７図は解

放を実施したあと更新した記憶装置管理テーブルを示した図である。

1…ユーザプログラム、2…ファイルシステム、3…入出力制御部、4…記憶装置、21…ファイル管理装置、22…データ解放制御部、23…データ割当制御部、24…仮想パーティション管理テーブル、25…記憶装置管理テーブル、231…記憶装置数・割り当て量算出手段、232…記憶装置選択手段。

代理人　弁理士　内　原　　晋

第1図



第2図

24 仮想パーティション管理テーブル

パーティション	VP
転送速度	T
必要格納量	U
要求処理速度	V
装置数	N

第3図

25 記憶装置管理テーブル

記憶装置ID	DK1	DK2	…	DKn
格納可能量	M1	M2	…	Mn
入出力待ち行列長	Q1	Q2	…	Qn
入出力要求量	L1	L2	…	Ln

第4図

24 仮想パーティション管理テーブル

パーティション	VP
転送速度	100
必要格納量	20
要求処理速度	150
装置数	4

第5図

25 記憶装置管理テーブル

記憶装置ID	DK1	DK2	DK3	DK4
格納可能量	600	480	550	500
入出力待ち行列長	1	3	2	4
入出力要求量	2	4	4	6

第6図

25 記憶装置管理テーブル

記憶装置ID	DK1	DK2	DK3	DK4
格納可能量	520	480	490	500
入出力待ち行列長	2	3	3	4
入出力要求量	6	4	7	6

第7図

25 記憶装置管理テーブル

記憶装置ID	DK1	DK2	DK3	DK4
格納可能量	520	540	490	540
入出力待ち行列長	2	3	3	4
入出力要求量	6	4	7	6